Dynafetch: Dynamic Prefetching for GPUs with UVM Support

Binong Kiri Bey
Department of Computer Science and Engineering
Indian Institute of Technology
Kanpur, India
Email: binong@cse.iitk.ac.in

Modern computing workloads using GPUs demand large memory capacity, usually exceeding the GPUs. The traditional GPU programming model of ”copy data to device, execute, fetch results to host” burdens programmers to split the large data into smaller chunks and explicitly copy data from Host-to-GPU or GPU-to-Host. Unified Virtual Memory (UVM) uses a unified virtual address space to effectively manage data in both the host’s and the GPU’s memory by using on-demand migration to avoid explicit copying. When data accessed by the GPU is not present in the GPU’s global memory, it is fetched from the memory of the Host (*far fault*). These *far faults* reduce the overall throughput. Prefetchers were introduced to reduce far faults by migrating additional pages along with the requested page by the GPU. This paper introduces Dynafetch, an adaptive prefetcher that changes the aggressiveness of prefetching in the runtime based on the workload’s far fault patterns.